Power Industry Dictionary for Text-Mining and Natural Language Processing Application

Proof of Concept

Karen Kim-Stevens, kkim@epri.com
EPRI Principal Project Manager, Radiation Safety

U.S. NRC Data Science and Artificial Intelligence
Regulatory Applications Workshops
Workshop #1 June 29, 2021

Today, NLP tools will parse words based on their more common usage

NRC.gov text

On April 6, 2006,
a drain noun
cooler noun
relief valve noun
in the
feedwater ? system noun
lifted and
remained open

Word indexing

A Generic NLP dictionary

drain

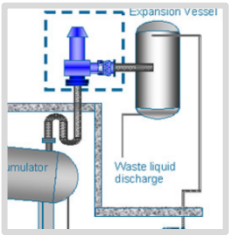
cooler

relief valve

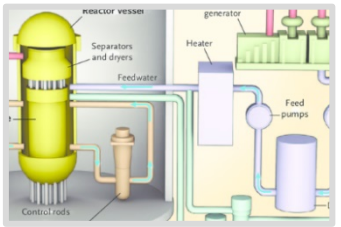
Vs.

B Groundwater Contamination Classifier

drain cooler
relief valve



feedwater system



Objective: Build a Nuclear Industry NLP Dictionary

Our *use case* for this proof of principle

Groundwater Contamination

Use case owner: Karen Kim-Stevens

Goal: Develop a NLP proof of principle that demonstrates the potential benefits of machine learning applied to this domain.

Tasks

- Create a preliminary dictionary to be used for classification.
- Develop a NLP text analytic demo and generate preliminary insights.

Benefits

Natural language text analytics will help the industry enhance preparation and implementation of mitigating actions in the event of inadvertent leaks and spills of radioactive materials.

Scenario

The industry has thousands of text documents from operating experiences, maintenance reports, work orders, regulatory filings, and more that reference Groundwater Contamination. Given the safety significance and the need to find ways to operate more efficiently, all nuclear plants would benefit from extracting and sharing key information from these documents to make quicker, informed decisions, reduce the number of inadvertent spills and leaks, and enhance the safety and response time to a contamination situation.

Potential use cases to develop risk mitigation strategies

Identify Specific SSCs

Identify which SSCs could be associated with a failure and release radioactive liquid into the environment

- How have the sources of SSC leaks and spills changed over time?
- Does the age of the plant impact the components?
- Do certain components leak after a certain amount of time in service?

Work Practices

Identify which work practice tasks could be associated to which jobs or systems that could cause the most release of radioactive liquid into the environment

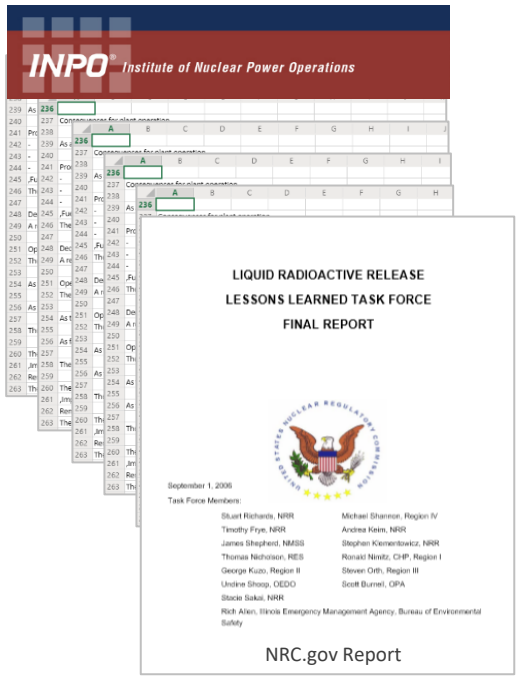
- Do work practices during planned vs. unplanned outages affect the prediction?
- Do routine vs. non-routine affect the prediction?
- How have leaks from work practices changed over time?

Concentration of radioactive material

Identify how concentration of radioactive material varies by type of leak or spill

- How much does the concentration vary by SSC or WP?
- Does the magnitude vary by for SSCs at BWR vs. PWR plants?
- Can this information be used to help plants identify the source of leaks or spills?

To ensure data integrity and quality results, we followed a structured data science approach



Identify Corpus

Develop list of initial words for dictionary

Review output with SME to validate

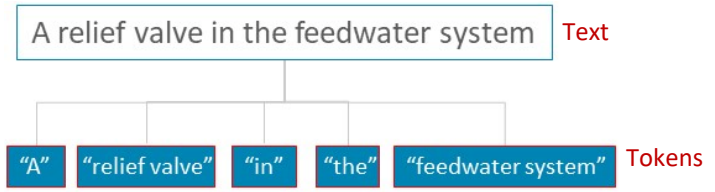
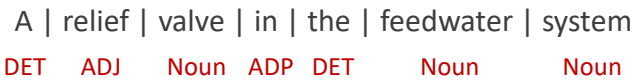
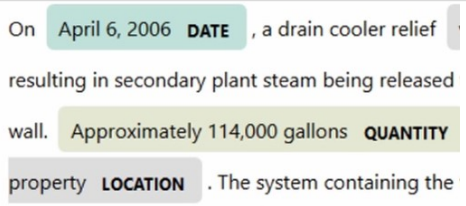
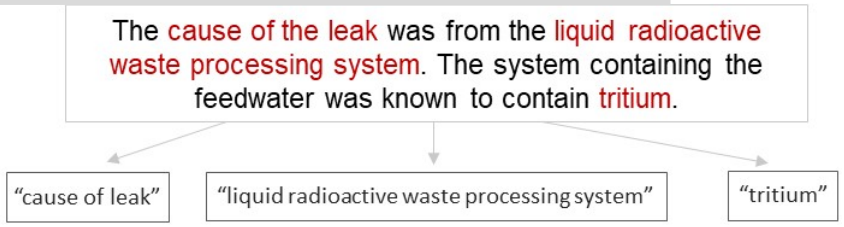
Export Dictionary

Develop expanded list of words

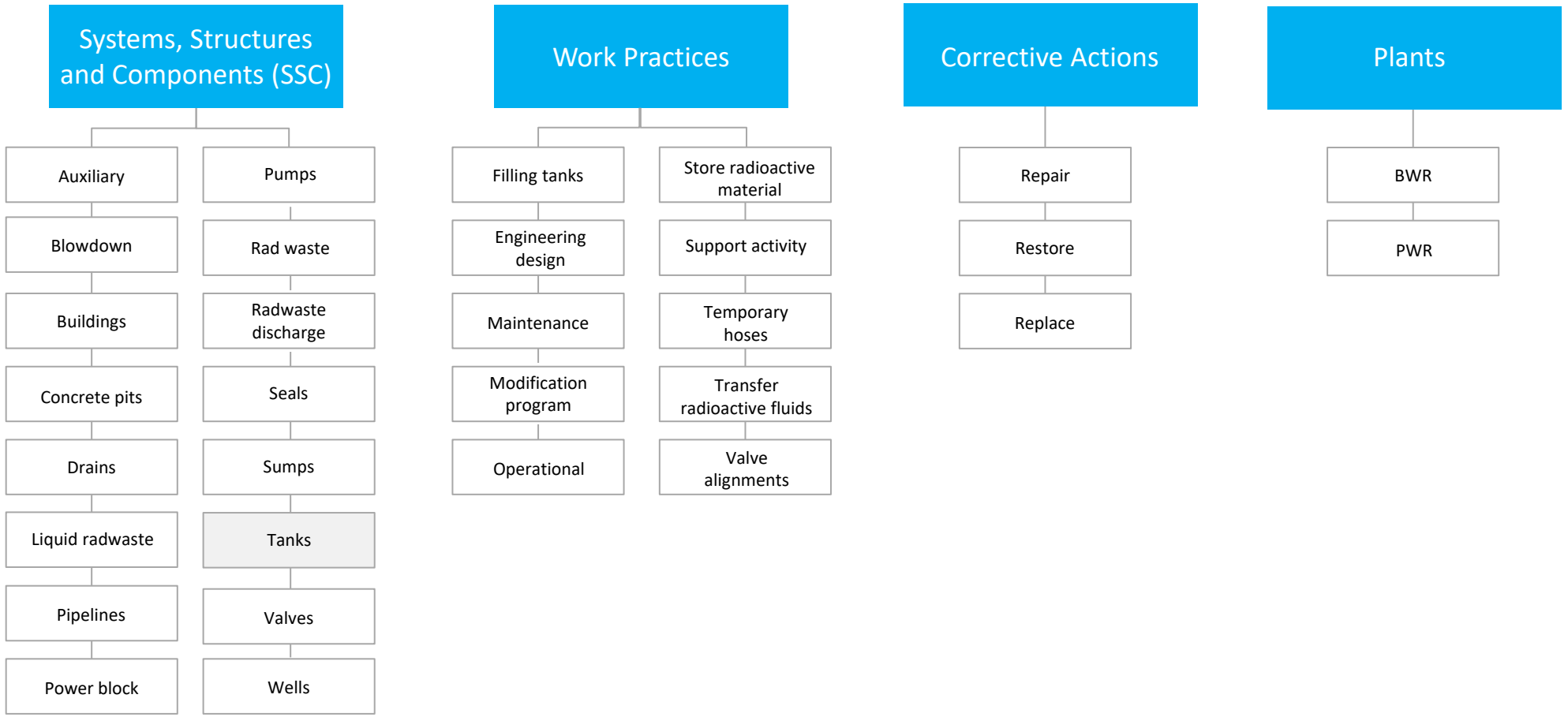
Create algorithms to search for additional words

Incorporate validated words into domain dictionary

These NLP techniques help to get preliminary results quickly

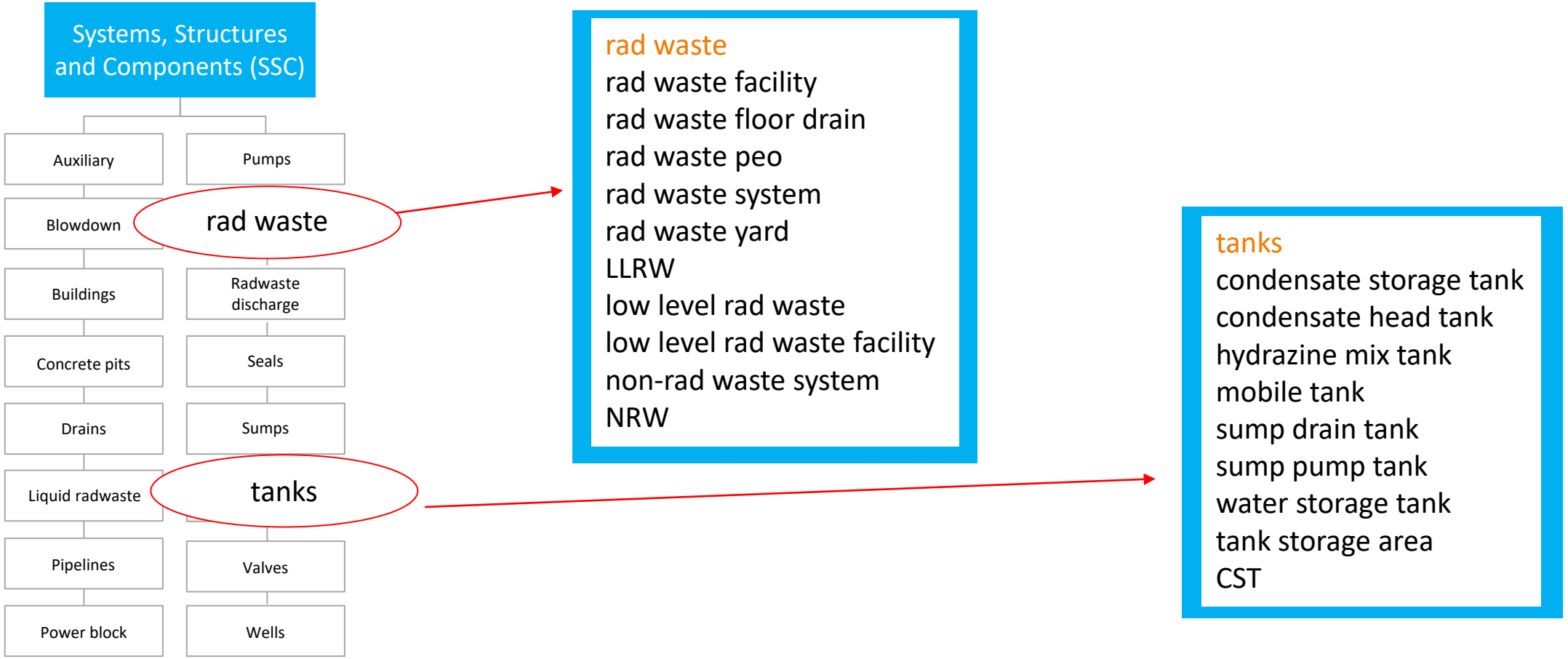
	What is it?	Why?	Example
1 Tokenization	Algorithms to segment text into groupings, phrases, punctuation, called “tokens.”	Tokens become the inputs for conducting text mining.	
2 Part of Speech	Statistical models that assign a part of speech to each token - noun, verb, adjective, adverb, etc.	These tags enable modeling to infer the relationships between words in phrases and sentences.	
3 Name Entity Recognition	Statistical models that assign labels to tokens such as date, quantity, location and more.	Entity recognition is helpful for information extraction and filtering.	
4 Rules-based Matching	Algorithms that find phrases, sequences of tokens, and entities.	Improves information extraction and text mining. This approach is one way to directly incorporate SME input.	

Architecture for the library of dictionaries



The dictionary map provides us with guidance on how topics are organized, overlap and relate to each other. The design evolves based on programmatic exploration and feedback from our SME.

Each word has an additional level of word associations that serve as training topics for machine learning



The lack of a consistent industry nomenclature is a key challenge in building NLP models

For example

groundwater
- gw
- ground water
- ground-water
- gnd water
- g water
- g-water

picocuries
- pCi/L
- pCi
- pCi / L
- picocuries / liter
- picocuries per liter
- pCi/liter
- pCi / liter

pits
- basins
- moats
- motes
- ponds

power block
- auxiliary building
- auxiliary system
- rad waste building
- radwaste building

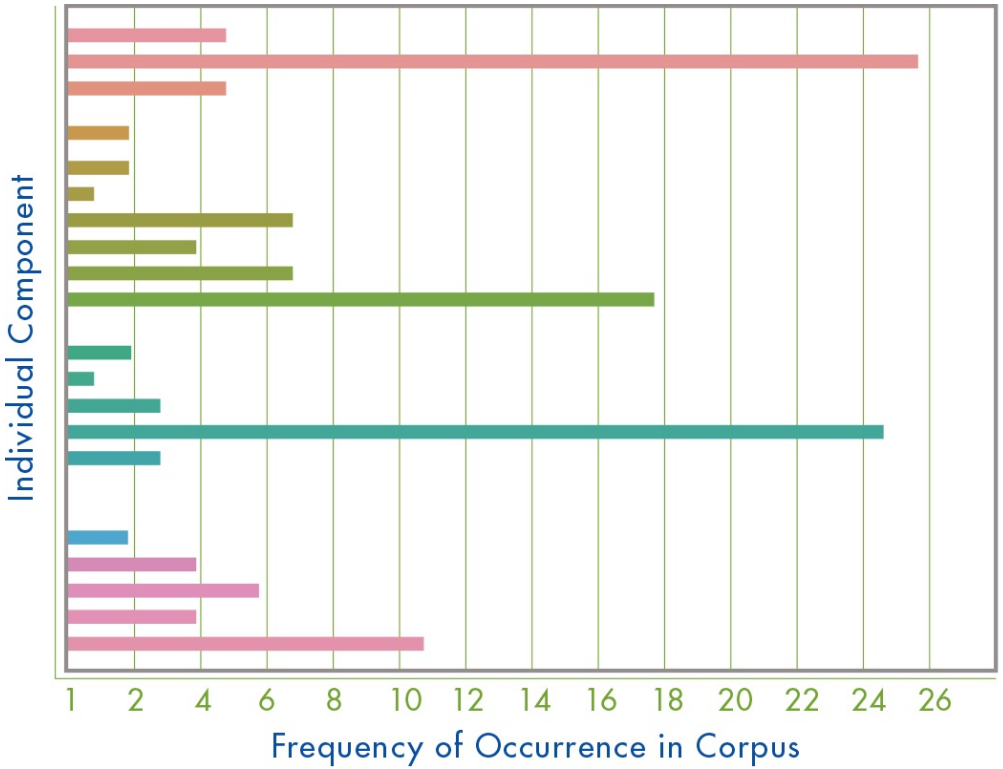
seismic gap
- cracks
- rattle space
- seals
- structural joints

storm drains
- drain systems
- roof drains
- storm systems
- yard drains

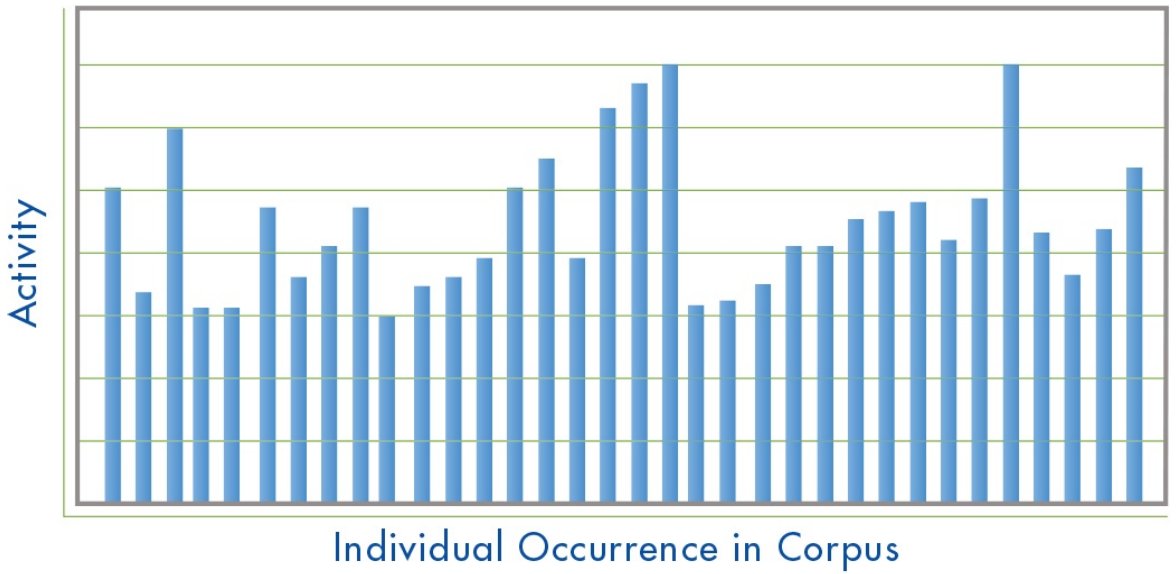
These word variabilities were identified through NLP algorithms and augmented by our subject matter expert (SME). A SME needs to provide guidance as we use the NLP tools.

Preliminary Visualizations

Counts for systems, structures, components associated with groundwater incident reports



Tritium activity (logarithmic scale) of reported events extracted from corpus for one type of reactor design



Elements of our work are transferable
and can help others get started on a similar project

Roles to make this type of project successful

- Project Sponsor
  - The project sponsor is the person or group who owns the project. They hold overall accountability of the project and are responsible for providing resources, support and guidance to enable success. This role ensures that the analysis is aligned with research and business goals.
- Subject Matter Expert (SME)
  - The SME plays a vital role in helping the data scientist understand the data and its nuances. This role will evaluate the text analytic output, ensure it is producing relevant results, and help to describe the specific real world problem that the machine learning project is trying to solve.
- Data Scientist
  - A data scientist collects, analyzes, and interprets large amounts of data. Their skills and expertise in highly advanced analytical tools enables them to understand the data and develop operational models, systems and tools by applying experimental and iterative methods and techniques.
- Data Analyst
  - A data analyst examines the patterns, trends, and other insights extracted from the data. They are responsible for deriving meaningful, actionable insights from the data. They support the project by creating visualizations.

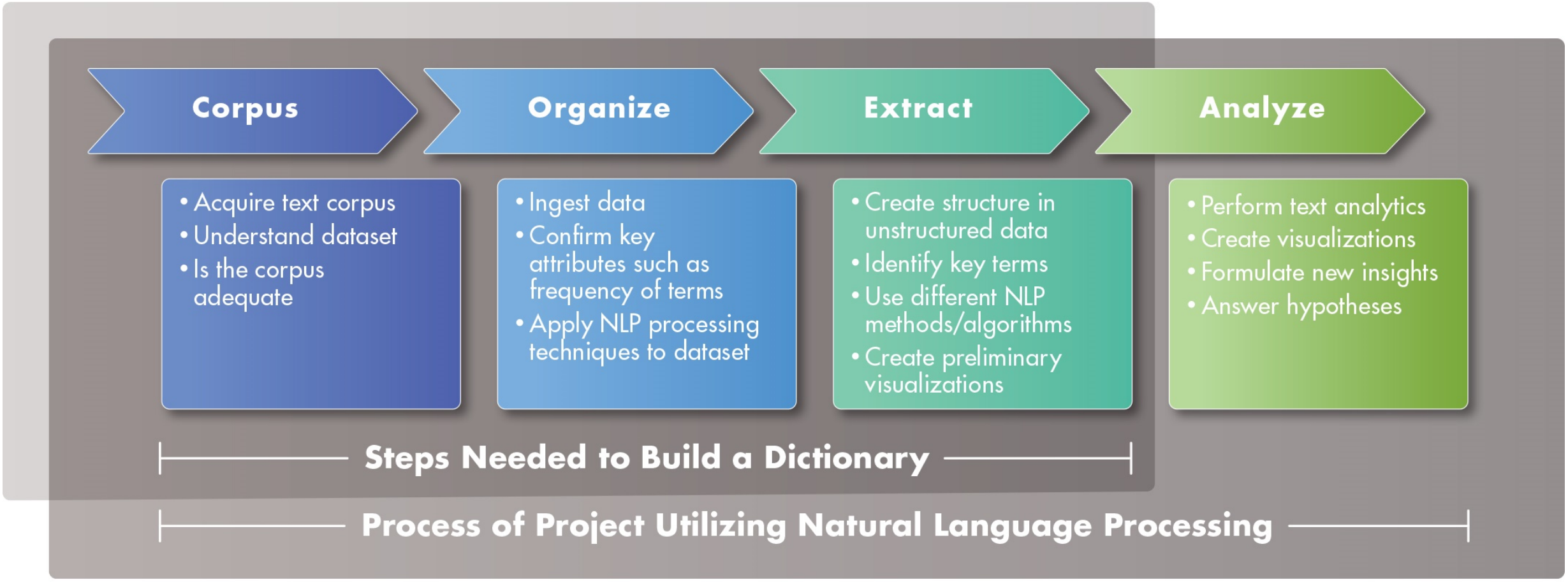
Key Takeaways

- Open-source dictionaries do not understand electric power industry language
- An industry-specific dictionary is needed to conduct text mining and apply NLP-based algorithm
- A workflow template for dictionary construction is repeatable that can be applied to new topics
- The development of an industry specific dictionary will require investment
- However, the nuclear industry will benefit from more efficient ways of digesting and applying industry data and knowledge

For More Information:

Please download “Quick Insight – Power Industry Dictionary for Text-Mining and Natural Language Processing: Proof of Concept.”
https://www.epri.com/research/products/000000003002019609

High Level Process for NLP Projects



High Level Process for NLP Projects

Corpus

- Acquire text corpus
- Understand dataset
- Is the corpus adequate

Organize

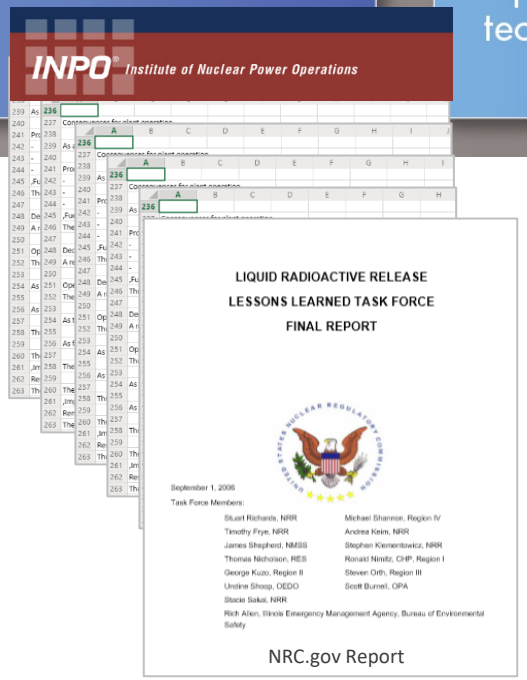
- Ingest data
- Confirm key attributes such as frequency of terms
- Apply NLP processing techniques to dataset

Extract

- Create structure in unstructured data
- Identify key terms
- Use different NLP methods/algorithms
- Create preliminary visualizations

Analyze

- Perform text analytics
- Create visualizations
- Formulate new insights
- Answer hypotheses



NRC.gov Report

High Level Process for NLP Projects

Corpus

- Acquire text corpus
- Understand dataset
- Is the corpus adequate

Organize

- Ingest data
- Confirm key attributes such as frequency of terms
- Apply NLP processing techniques to dataset

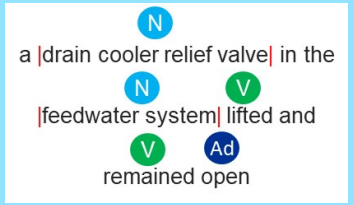
Extract

- Create structure in unstructured data
- Identify key terms
- Use different NLP methods/algorithms
- Create preliminary visualizations

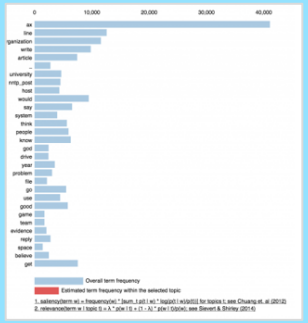
Analyze

- Perform text analytics
- Create visualizations
- Formulate new insights
- Answer hypotheses

Parts-of-Speech Tagging

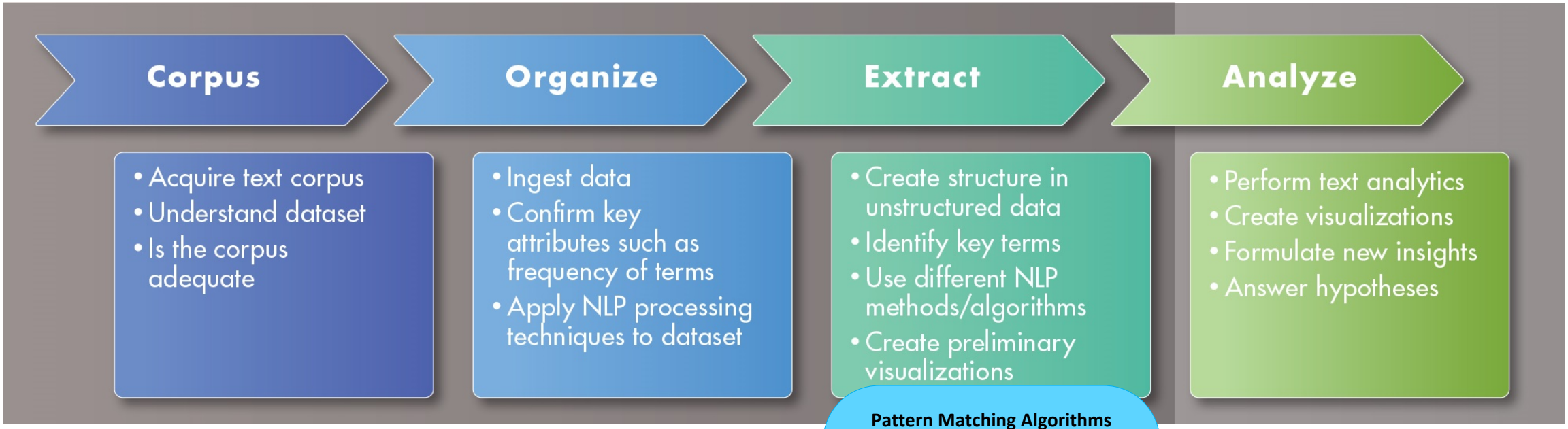


Word Frequency



Move forward, when you have enough understanding of the text to work with it as data

High Level Process for NLP Projects



Move forward, when you have enough structure to begin text mining

High Level Process for NLP Projects

Corpus

- Acquire text corpus
- Understand dataset
- Is the corpus adequate

Organize

- Ingest data
- Confirm key attributes such as frequency of terms
- Apply NLP processing techniques to dataset

Extract

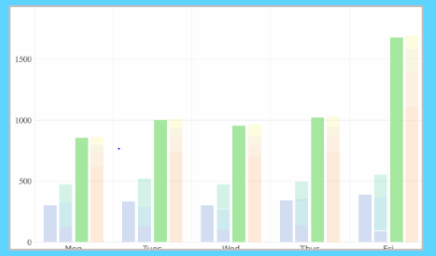
- Create structure in unstructured data
- Identify key terms
- Use different NLP methods/algorithms
- Create preliminary visualizations

Analyze

- Perform text analytics
- Create visualizations
- Formulate new insights
- Answer hypotheses

Iterate. The analysis will determine if more extractions are needed

Magnitude of Leaks/Spills



High Risk Components PWR plants